Pathway Fingerprinting - example script

Gabriel Altschuler

December 8, 2011

This document demonstrates how to use the `pathprint` package to analyze a dataset using Pathway Fingerprints. The pathprint package takes gene expression data and processes this into discrete expression scores (+1,0,-1) for a set of 633 pathways. For more information, see the pathprint website.

Contents

1 Initial data processing

An existing GEO sample on the Human Affy ST 1.0 chip will be used as en example. The dataset GSE26946 profiles expression data from iPS and human ES cells. The R package `GEOquery` can be used to retrieve the data. An 'exprs' object, i.e. a dataframe with row names corresponding to probe or feature IDs and column names corresponding to sample IDs is required by `pathprint`. In addition, we need to know the GEO reference for the platform, in this case GPL6244, and the species, which is 'human' or "Homo sapiens' (both styles of name work).

```
> library(GEOquery)
> GSE26946 <- getGEO("GSE26946")
> GSE26946.exprs <- exprs(GSE26946[[1]])
> GSE26946.exprs[1:5, 1:3]
```

```
        GSM663450 GSM663451 GSM663452
7892501  8.904383  9.328561  8.760057
7892502  7.217361  9.118137  6.242542
7892503  6.091620  5.620844  5.726464
7892504 11.072690 10.883280 10.714790
7892505  5.777377  4.814570  4.463360
```

```
> GSE26946.platform <- annotation(GSE26946[[1]])
> GSE26946.species <- as.character(unique(phenoData(GSE26946[[1]])$organism_ch1))
> GSE26946.names <- as.character(phenoData(GSE26946[[1]])$title)
```

2 Pathway fingerprinting

2.1 Fingerprinting from new expression data

Now the data has been prepared, the `pathprint` function `exprs2fingerprint` can be used to produce a pathway fingerprint from this expression table.

```
> library(pathprint)
> GSE26946.fingerprint <- exprs2fingerprint(exprs = GSE26946.exprs,
                                            platform = GSE26946.platform,
                                            species = GSE26946.species,
                                            progressBar = TRUE
                                            )
```

```
[1] "Running fingerprint"
```

```
> GSE26946.fingerprint[1:5, 1:3]
```

```
                                                GSM663450
Glycolysis / Gluconeogenesis (KEGG)                     1
Citrate cycle (TCA cycle) (KEGG)                        1
Pentose phosphate pathway (KEGG)                        1
Pentose and glucuronate interconversions (KEGG)         1
Fructose and mannose metabolism (KEGG)                  1
                                                GSM663451
Glycolysis / Gluconeogenesis (KEGG)                     1
Citrate cycle (TCA cycle) (KEGG)                        1
Pentose phosphate pathway (KEGG)                        1
Pentose and glucuronate interconversions (KEGG)         1
Fructose and mannose metabolism (KEGG)                  1
                                                GSM663452
Glycolysis / Gluconeogenesis (KEGG)                     1
Citrate cycle (TCA cycle) (KEGG)                        1
Pentose phosphate pathway (KEGG)                        1
Pentose and glucuronate interconversions (KEGG)         1
Fructose and mannose metabolism (KEGG)                  1
```

2.2 Using existing data

The pathprint package contains the object `GEO.fingerprint.matrix` which contains 188390 samples that have already been fingerprinted, along with their associated metadata, in the object `GEO.metadata.matrix`. As the above data record is publically available from GEO it is actually already in the matrix and we can compare this to the fingerprint processed above. It should be noted that occasionally there may be discrepancies in one or two pathways due to the way in which the threshold is applied. Any existing GEO record not contained within the fingerprint matrix must either a) be on a chip that is not covered or b) have been uploaded to GEO in a not standard manner, normally with an incorrectly matched species name or an unusually small number or probes or feature IDs.

```
> colnames(GSE26946.exprs) %in% colnames(GEO.fingerprint.matrix)
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

> GSE26946.existing <- GEO.fingerprint.matrix[,colnames(GSE26946.exprs)]
> all.equal(GSE26946.existing, GSE26946.fingerprint)

[1] TRUE
```

3 Fingerprint Analysis

3.1 Intra-sample comparisons

The fingerprint vectors can be used to compare the differntially expressed functions within the sample set. The most straight forward method to represent this is using a heatmap, removing rows for which there is no change in functional expression.

```
> heatmap(GSE26946.fingerprint[apply(GSE26946.fingerprint, 1, sd) > 0, ],
         labCol = GSE26946.names,
         mar = c(10,10),
         col = c("blue", "white", "red"))
```

3.2 Using consensusFingerprint and fingerprinDistance, comparison to pluripotent arrays

We can also investigate how far in functional distance, these arrays are from other pluripotent fingerprints. This can be achieved using the set of pluripotent arrays included in the package, from which a consensus fingerprint can be created.

```
> # construct pluripotent consensus
> pluripotent.consensus<-consensusFingerprint(
   GEO.fingerprint.matrix[,pluripotents.frame$GSM], threshold=0.9)
> # calculate distance from the pluripotent consensus for all arrays
```

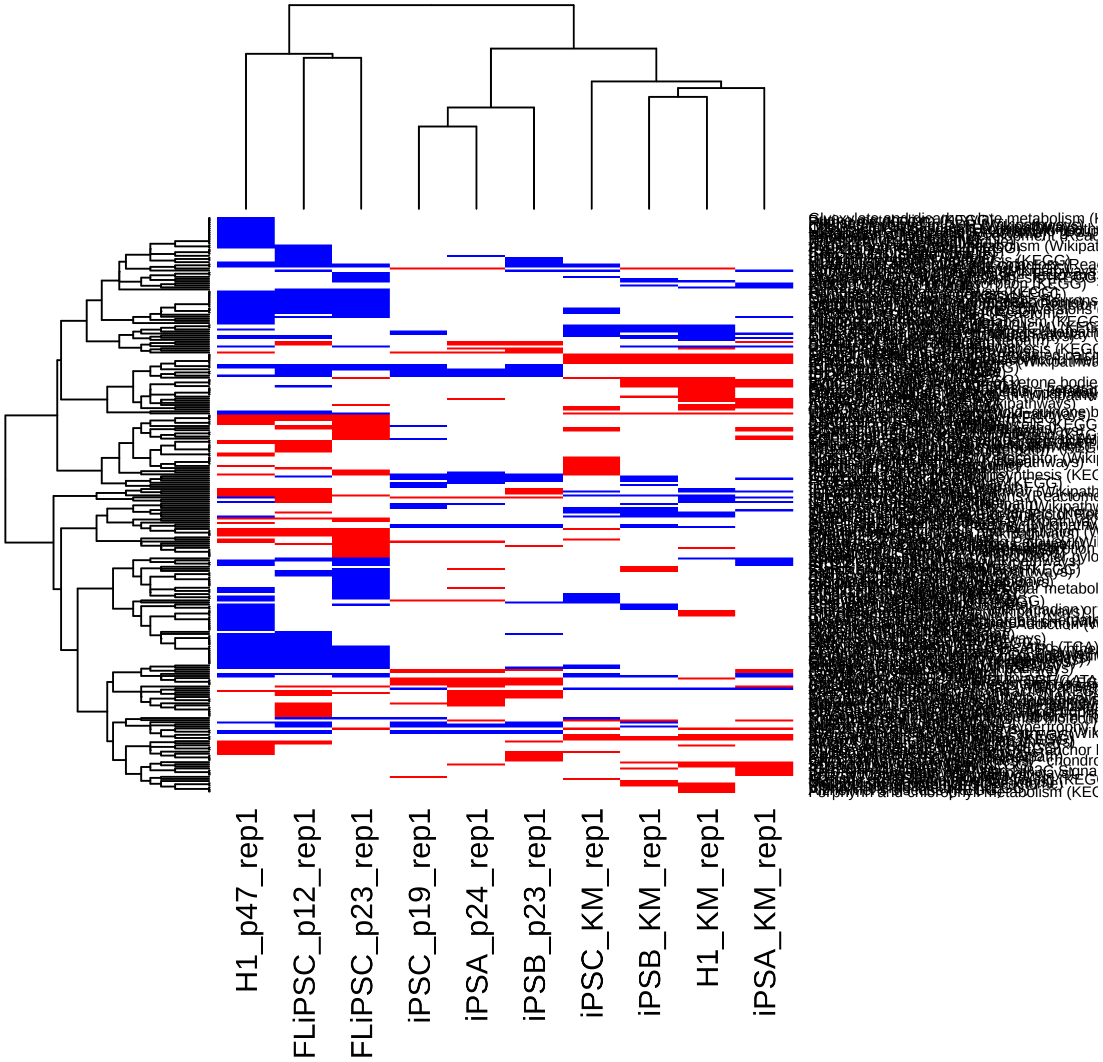

Figure 1: Heatmap of GSE26946 pathway fingerprints, blue = -1, white = 0, red = +1

```
> geo.pluripotentDistance<-consensusDistance(
         pluripotent.consensus, GEO.fingerprint.matrix)
[1] "Scaling against max length, 180"
> # calculate distance from pluripotent consensus for GSE26946 arrays
> GSE26946.pluripotentDistance<-consensusDistance(
   pluripotent.consensus, GSE26946.fingerprint)
[1] "Scaling against max length, 180"
```

```
> par(mfcol = c(2,1), mar = c(0, 4, 4, 2))
> geo.pluripotentDistance.hist<-hist(geo.pluripotentDistance[,"distance"],
         nclass = 50, xlim = c(0,1), main = "Distance from pluripotent consensus")
> par(mar = c(7, 4, 4, 2))
> hist(geo.pluripotentDistance[pluripotents.frame$GSM, "distance"],
         breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
         main = "", xlab = "")
> hist(GSE26946.pluripotentDistance[, "distance"],
   breaks = geo.pluripotentDistance.hist$breaks, xlim = c(0,1),
         main = "", col = "red", add = TRUE)
```

3.3 Identifying similar arrays

We can use the data contained within the GEO fingerprint matrix to order all of the GEO records according to distance from an experiment (or set of experiments, see below). This can be used, in conjunction with the metadata, to annotate a fingerprint with data from the GEO corpus. Here, we will identify experiments closely matched to the H1, embyonic stem cells within GSE26946

```
> GSE26946.H1<-consensusFingerprint(
   GSE26946.fingerprint[,grep("H1", GSE26946.names)], threshold=0.9)
> geo.H1Distance<-consensusDistance(
   GSE26946.H1, GEO.fingerprint.matrix)
[1] "Scaling against max length, 700"
> # look at top 20
> GEO.metadata.matrix[
   match(head(rownames(geo.H1Distance),20), GEO.metadata.matrix$GSM),
   c("GSM", "GSE", "GPL", "Source")]
             GSM      GSE     GPL
160351 GSM663458 GSE26946 GPL6244
160352 GSM663459 GSE26946 GPL6244
160348 GSM663455 GSE26946 GPL6244
160346 GSM663453 GSE26946 GPL6244
160344 GSM663451 GSE26946 GPL6244
160350 GSM663457 GSE26946 GPL6244
160347 GSM663454 GSE26946 GPL6244
160345 GSM663452 GSE26946 GPL6244
129376 GSM525412 GSE21037 GPL6244
```

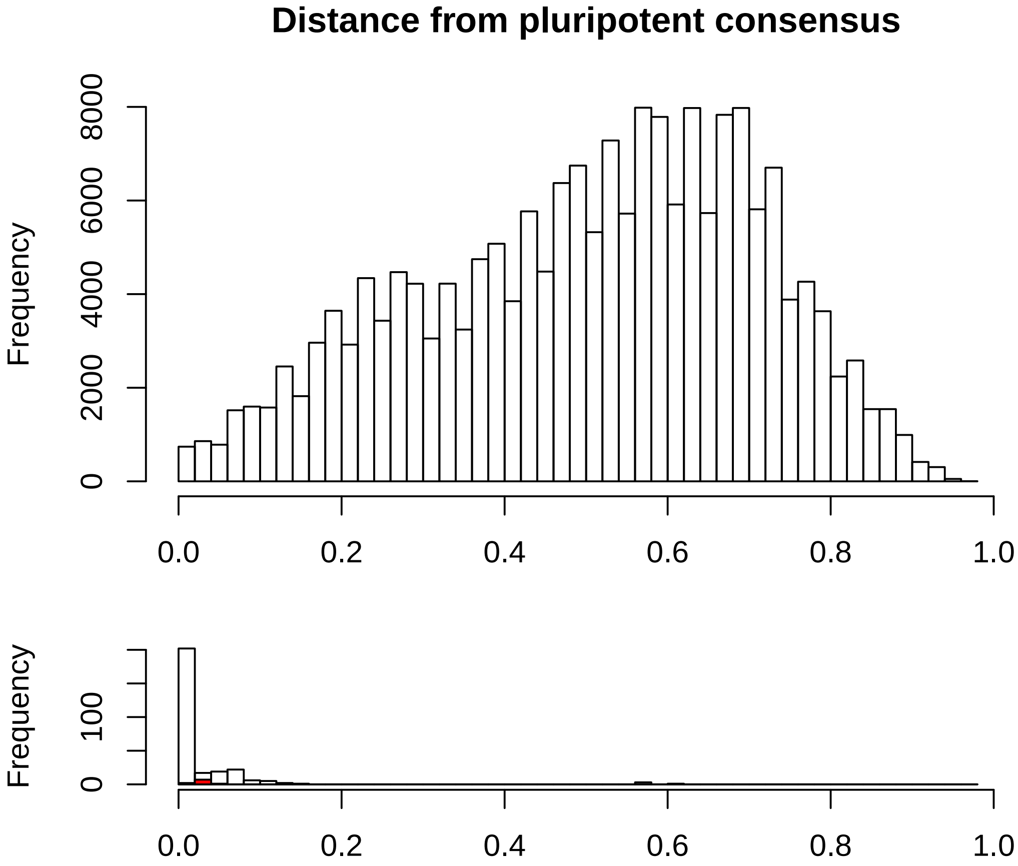

Figure 2: Histogram representing the distance from the pluripotent consensus fingerprint for all GEO (above), curated pluripotent samples (below), and GSE26946 samples (below, red)

```
160343 GSM663450 GSE26946 GPL6244
160349 GSM663456 GSE26946 GPL6244
129374 GSM525410 GSE21037 GPL6244
129373 GSM525409 GSE21037 GPL6244
129375 GSM525411 GSE21037 GPL6244
165300 GSM697683 GSE21655 GPL6244
129378 GSM525414 GSE21037 GPL6244
165299 GSM697682 GSE21655 GPL6244
129377 GSM525413 GSE21037 GPL6244
129385 GSM525421 GSE21037 GPL6244
129384 GSM525420 GSE21037 GPL6244
                                        Source
160351              Human Embryonic Stem Cells
160352              Human Embryonic Stem Cells
160348          induced pluripotent stem cells
160346          induced pluripotent stem cells
160344          induced pluripotent stem cells
160350          induced pluripotent stem cells
160347          induced pluripotent stem cells
160345          induced pluripotent stem cells
129376                       iPSC-RTT clone 18
160343          induced pluripotent stem cells
160349          induced pluripotent stem cells
129374                       iPSC-RTT clone 15
129373                       iPSC-RTT clone 15
129375                       iPSC-RTT clone 15
165300 iPSC maintained under feeder conditions
129378                       iPSC-RTT clone 18
165299 iPSC maintained under feeder conditions
129377                       iPSC-RTT clone 18
129385                         iPSC-WT clone 2
129384                         iPSC-WT clone 1
```